EXPRESS MAIL NO. EL 634621280US

PA-0038US

GENES EXPRESSED IN COLON CANCER

FIELD OF THE INVENTION

The present invention relates to a combination comprising a plurality of cDNAs which are differentially expressed in colon cancer and in premalignant conditions of the colon and which may be used entirely or in part to diagnose, to stage, to treat, or to monitor the progression or treatment of colon cancer.

BACKGROUND OF THE INVENTION

Array technology can provide a simple way to explore the expression of a single polymorphic gene or the expression profile of a large number of related or unrelated genes. When the expression of a single gene is examined, arrays are employed to detect the expression of a specific gene or its variants. When an expression profile is examined, arrays provide a platform for examining which genes are tissue specific, carrying out housekeeping functions, parts of a signaling cascade, or specifically related to a particular genetic predisposition, condition, disease, or disorder.

The potential application of gene expression profiling is particularly relevant to improving diagnosis, prognosis, and treatment of disease. For example, both the levels and sequences expressed in tissues from subjects with colon cancer may be compared with the levels and sequences expressed in normal tissue.

Colorectal cancer is the fourth most common cancer and the second most common cause of cancer death in the United States with approximately 130,000 new cases and 55,000 deaths per year. Colon and rectal cancers share many environmental risk factors and both are found in individuals with specific genetic syndromes. (See Potter (1999) J Natl Cancer Institute 91:916-932 for a review of colorectal cancer.) Colon cancer is the only cancer that occurs with approximately equal frequency in men and women, and the five-year survival rate following diagnosis of colon cancer is around 55% in the United States (Ries et al. (1990) National Institutes of Health, DHHS Publ No. (NIH)90-2789).

Colon cancer is causally related to both genes and the environment. Several molecular pathways have been linked to the development of colon cancer, and the expression of key genes in any of these pathways may be lost by inherited or acquired mutation or by hypermethylation. There is a particular need to identify genes for which changes in expression may provide an early indicator of colon cancer or a predisposition for the development of colon cancer.

For example, it is well known that abnormal patterns of DNA methylation occur consistently in human tumors and include, simultaneously, widespread genomic hypomethylation and localized areas of increased methylation. In colon cancer in particular, it has been found that these changes occur early in tumor progression such as in premalignant polyps that precede colon cancer. Indeed, DNA methyltransferase, the enzyme that performs DNA methylation, is significantly increased in histologically normal mucosa from patients with colon cancer or the benign polyps that precede

cancer, and this increase continues during the progression of colonic neoplasms (Wafik et al. (1991) Proc Natl Acad Sci USA 88:3470-3474). Increased DNA methylation occurs in G+C rich areas of genomic DNA termed "CpG islands" that are important for maintenance of an "open" transcriptional conformation around genes, and that hypermethylation of these regions results in a "closed" conformation that silences gene transcription. It has been suggested that the silencing or downregulation of differentiation genes by such abnormal methylation of CpG islands may prevent differentiation in immortalized cells (Anteguera et al. (1990) Cell 62:503-514).

Familial Adenomatous Polyposis (FAP) is a rare autosomal dominant syndrome that precedes colon cancer and is caused by an inherited mutation in the adenomatous polyposis coli (APC) gene. FAP is characterized by the early development of multiple colorectal adenomas that progress to cancer at a mean age of 44 years. The APC gene is a part of the APC-ß-catenin-Tcf (T-cell factor) pathway. Impairment of this pathway results in the loss of orderly replication, adhesion, and migration of colonic epithelial cells that results in the growth of polyps. A series of other genetic changes follow activation of the APC-ß-catenin-Tcf pathway and accompanies the transition from normal colonic mucosa to metastatic carcinoma. These changes include mutation of the K-Ras proto-oncogene, changes in methylation patterns, and mutation or loss of the tumor suppressor genes p53 and Smad4/ DPC4. While the inheritance of a mutated APC gene is a rare event, the loss or mutation of APC and the consequent effects on the APC-ß-catenin-Tcf pathway is believed to be central to the majority of colon cancers in the general population.

Hereditary nonpolyposis colorectal cancer (HNPCC) is another inherited autosomal dominant syndrome with a less well defined phenotype than FAP. HNPCC, which accounts for about 2% of colorectal cancer cases, is distinguished by the tendency to early onset of cancer and the development of other cancers, particularly those involving the endometrium, urinary tract, stomach and biliary system. HNPCC results from the mutation of one or more genes in the DNA mis-match repair (MMR) pathway. Mutations in two human MMR genes, MSH2 and MLH1, are found in a large majority of HNPCC families identified to date. The DNA MMR pathway identifies and repairs errors that result from the activity of DNA polymerase during replication. Furthermore, loss of MMR activity contributes to cancer progression through accumulation of other gene mutations and deletions, such as loss of the BAX gene which controls apoptosis, and the TGFß receptor II gene which controls cell growth. Because of the potential for irreparable damage to DNA in an individual with a DNA MMR defect, progression to carcinoma is more rapid than usual.

Although ulcerative colitis is a minor contributor to colon cancer, affected individuals have about a 20-fold increase in risk for developing cancer. Progression is characterized by loss of the p53 gene which may occur early, appearing even in histologically normal tissue. The progression of the disease from ulcerative colitis to dysplasia/carcinoma without an intermediate polyp state suggests a

high degree of mutagenic activity resulting from the exposure of proliferating cells in the colonic mucosa to the colonic contents.

Almost all colon cancers arise from cells in which the estrogen receptor (ER) gene has been silenced. The silencing of ER gene transcription is age related and linked to hypermethylation of the ER gene (Issa et al. (1994) Nature Genetics 7:536-540). Introduction of an exogenous ER gene into cultured colon carcinoma cells results in marked growth suppression. The connection between loss of the ER protein in colonic epithelial cells and the consequent development of cancer has not been established.

Clearly there are a number of genetic alterations associated with colon cancer and with the development and progression of the disease, particularly the downregulation or deletion of genes, that potentially provide early indicators of cancer development, and which may also be used to monitor disease progression or provide possible therapeutic targets. The specific genes affected in a given case of colon cancer depend on the molecular progression of the disease. Identification of additional genes associated with colon cancer and the precancerous state would provide more reliable diagnostic patterns associated with the development and progression of the disease.

The present invention provides for a composition comprising a plurality of cDNAs for use in detecting changes in expression of genes encoding proteins associated with colon cancer. Such a composition satisfies a need in the art by providing a set of differentially expressed genes which may be used entirely or in part in the diagnosis, prognosis or treatment of colon cancer.

SUMMARY

The present invention provides a combination comprising a plurality of cDNAs and their complements which are differentially expressed in precancerous colon polyps and colon cancer and which are selected from SEQ ID NOs:1-3, 5, 6, 8-10,12, 14, 15, 17, 18, 20, 22, 24, 26-29, 31, 33, 34, 36-39, 41-43, 45-47, 49, 51, 53. 55-58, 60, 62, 64, 66, 67, 69, 71, 72, 74-79, 81, 83-86, 88, 89, 91, 92, 94, 96, 97, 99, 100, 102-104, 106, 107, 109, 111, 112, 114, 116, 118, 119, 121, 123-126, 128, 130, 131-137, 139, 140, 142-151, 153-157, 159, 160, 162-165, 167-172, 174, 176, 177, 179-181, 183-187, 189-191, and 193 as presented in the Sequence Listing. In one aspect, the combination is useful to diagnose a precancerous or cancerous condition in colon. In another aspect, the combination is immobilized on a substrate.

The invention also provides a combination comprising a subset of these cDNAs and their complements which are differentially expressed in colon cancer relative to colon polyps or normal colon tissue and which are selected from SEQ ID NOs:172, 174, 176, 177, 179-181, 183-187, 189-191, and 193. In one aspect, the combination is useful to diagnose a colon cancer or the progression of a colon disorder from colon polyps to colon cancer.

The invention further provides a high throughput method to detect differential expression of

one or more of the cDNAs of the combination. The method comprises hybridizing the substrate comprising the combination with the nucleic acids of a sample, thereby forming one or more hybridization complexes, detecting the hybridization complexes, and comparing the hybridization complexes with those of a standard, wherein differences in the size and signal intensity of each hybridization complex indicates differential expression of nucleic acids in the sample. In one aspect, the sample is from a subject with colon cancer and differential expression determines an early, mid, and late stage of that disorder.

The invention further provides a high throughput method of screening a library or plurality of molecules or compounds to identify a ligand. The method comprises combining the substrate comprising the combination with a library or plurality of molecules or compounds under conditions to allow specific binding and detecting specific binding, thereby identifying a ligand. The library or plurality of molecules or compounds are selected from DNA molecules, RNA molecules, peptide nucleic acid molecules, mimetics, peptides, transcription factors, repressors, and other regulatory proteins. The invention additionally provides a method for purifying a ligand, the method comprising combining a cDNA of the invention with a sample under conditions which allow specific binding, recovering the bound cDNA, and separating the cDNA from the ligand, thereby obtaining purified ligand.

The invention still further provides an isolated cDNA selected from SEQ ID NOs:12, 41, 71, 74, 154,162, 167, 170, and 177 as presented in the Sequence Listing. The invention also provides a vector comprising the cDNA, a host cell comprising the vector, and a method for producing a protein comprising culturing the host cell under conditions for the expression of a protein and recovering the protein from the host cell culture.

The present invention provides a purified protein encoded and produced by a cDNA of the invention. The invention also provides a high-throughput method for using a protein to screen a library or a plurality of molecules or compounds to identify a ligand. The method comprises combining the protein or a portion thereof with the library or plurality of molecules or compounds under conditions to allow specific binding and detecting specific binding, thereby identifying a ligand which specifically binds the protein. A library or plurality of molecules or compounds are selected from DNA molecules, RNA molecules, peptide nucleic acid molecules, mimetics, peptides, proteins, agonists, antagonists, antibodies or their fragments, immunoglobulins, inhibitors, drug compounds, and pharmaceutical agents. The invention further provides for using a protein to purify a ligand. The method comprises combining the protein or a portion thereof with a sample under conditions to allow specific binding, recovering the bound protein, and separating the protein from the ligand, thereby obtaining purified ligand. The invention still further provides a composition comprising the protein and a pharmaceutical carrier.

The invention also provides methods for using a protein to prepare and purify polyclonal and monoclonal antibodies which specifically bind the protein. The method for preparing a polyclonal antibody comprises immunizing a animal with protein under conditions to elicit an antibody response, isolating animal antibodies, attaching the protein to a substrate, contacting the substrate with isolated antibodies under conditions to allow specific binding to the protein, dissociating the antibodies from the protein, thereby obtaining purified polyclonal antibodies. The method for preparing and purifying monoclonal antibodies comprises immunizing a animal with a protein under conditions to elicit an antibody response, isolating antibody producing cells from the animal, fusing the antibody producing cells with immortalized cells in culture to form monoclonal antibody producing hybridoma cells, culturing the hybridoma cells, and isolating from culture monoclonal antibodies which specifically bind the protein.

The invention provides a purified antibody that specifically binds a protein expressed in colon cancer. The invention also provides a method for using an antibody to detect expression of a protein in a sample comprising combining the antibody with a sample under conditions which allow the formation of antibody:protein complexes and detecting complex formation, wherein complex formation indicates expression of the protein in the sample.

DESCRIPTION OF THE SEQUENCE LISTING AND TABLES

A portion of the disclosure of this patent document contains material that is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

The Sequence Listing is a compilation of cDNAs obtained by sequencing and extension of clone inserts. Each sequence is identified by a sequence identification number (SEQ ID NO) and by the template number (TEMPLATE ID) from which it was obtained.

Table 1 lists the differential expression of clones representing the cDNAs of the present invention that are differentially expressed in both colon polyps and colon cancer relative to normal colon tissue. Column 1 lists the Incyte cDNA Clone ID, and columns 2-11 list the differential expression values observed in colon tissue samples from patients with colon polyps (columns 2-4) and colon cancer (columns 5-11).

Table 2 lists the differential expression values of clones representing cDNAs that are differentially expressed in colon polyps and colon cancer relative to normal colon in which the expression in colon cancer is found to be statistically more significant than in colon polyps. Column 1 lists the Incyte Clone ID, columns 2-9 list the differential expression values in colon polyps (columns 2-4) and colon cancer (columns 5-9), and column 10 lists the value of the student t-test for the significance between the expression in colon polyps versus colon cancer.

Table 3 links the differentially expressed clones on a microarray with Incyte cDNA templates. Columns 1 and 2 show the SEQ ID NO and TEMPLATE ID, respectively. Column 3 shows the CLONE ID and columns 4 and 5 show the first residue (START) and last residue (STOP) encompassed by the clone on the template.

Table 4 shows Incyte nucleotide templates presented in the Sequence Listing and the corresponding protein templates encoded by these cDNAs, also presented in the sequence Listing. Columns 1 and 2 show the SEQ ID NO and the Nucleotide Template ID, respectively, and columns 3 and 4 show the corresponding SEQ ID NO and Protein Template ID, respectively.

Table 5 shows the annotation of both nucleotide and protein Template IDs of the invention to sequences in GenBank. Columns 1 and 2 show the SEQ ID NO and Template ID, respectively. Columns 3, 4, and 5 show the GenBank hit (GI Number), probability score (E-value), and functional annotation, respectively, as determined by BLAST analysis (version 1.4 using default parameters; Altschul (1993) J Mol Evol 36: 290-300; Altschul et al. (1990) J Mol Biol 215:403-410) of the cDNA against GenBank (release 116; National Center for Biotechnology Information (NCBI), Bethesda MD).

Table 6 shows Pfam annotations of the cDNAs and proteins of the present invention. Columns 1 and 2 show the SEQ ID NO and TEMPLATE ID, respectively. Columns 3 and 4 show the first residue (START), last residue (STOP), respectively, for the segment of the cDNA or protein identified by Pfam analysis. Column 5 shows the reading frame for cDNA sequences. Columns 6 and 7 show the Pfam hit and Pfam description, respectively, corresponding to the polypeptide domain encoded by the cDNA segment or found in the protein sequence, and column 8 shows the E-value for the annotation.

Table 7 shows signal peptide and transmembrane regions predicted within the cDNAs of the present invention and in the proteins of the invention. Columns 1 and 2 show the SEQ ID NO and TEMPLATE ID, respectively. Columns 3 and 4 show the first residue (START), last residue (STOP), respectively, for the segment of the cDNA or the protein identified as a signal peptide or transmembrane region, and column 5 shows the reading frame for cDNA sequences. Column 6 identifies the polypeptide region as either a signal peptide (SP) or transmembrane (TM) domain.

DESCRIPTION OF THE INVENTION

Definitions

"Array" refers to an ordered arrangement of at least two cDNAs on a substrate. At least one of the cDNAs represents a control or standard sequence, and the other, a cDNA of diagnostic interest. The arrangement of from about two to about 40,000 cDNAs on the substrate assures that the size and signal intensity of each labeled hybridization complex formed between a cDNA and a

sample nucleic acid is individually distinguishable.

The "complement" of a nucleic acid molecule of the Sequence Listing refers to a nucleotide sequence which is completely complementary over the full length of the sequence and which will hybridize to the nucleic acid molecule under conditions of high stringency.

"cDNA" refers to a chain of nucleotides, an isolated polynucleotide, nucleic acid molecule, or any fragment or complement thereof. It may have originated recombinantly or synthetically, be double-stranded or single-stranded, coding and/or noncoding, an exon with or without an intron from a genomic DNA molecule, and purified or combined with carbohydrate, lipids, protein or inorganic elements or substances. Preferably, the cDNA is from about 400 to about 10,000 nucleotides.

The phrase "cDNA encoding a protein" refers to a nucleic acid sequence that closely aligns with sequences which encode conserved regions, motifs or domains that were identified by employing analyses well known in the art. These analyses include BLAST (Basic Local Alignment Search Tool; Altschul (1993) J Mol Evol 36: 290-300; Altschul et al. (1990) J Mol Biol 215:403-410) which provides identity within the conserved region. Brenner et al. (1998; Proc Natl Acad Sci 95:6073-6078) who analyzed BLAST for its ability to identify structural homologs by sequence identity found 30% identity is a reliable threshold for sequence alignments of at least 150 residues and 40% is a reasonable threshold for alignments of at least 70 residues (Brenner et al., page 6076, column 2).

"Derivative" refers to a cDNA or a protein that has been subjected to a chemical modification. Derivatization of a cDNA can involve substitution of a nontraditional base such as queosine or of an analog such as hypoxanthine. These substitutions are well known in the art. Derivatization of a protein involves the replacement of a hydrogen by an acetyl, acyl, alkyl, amino, formyl, or morpholino group. Derivative molecules retain the biological activities of the naturally occurring molecules but may confer advantages such as longer lifespan or enhanced activity.

"Differential expression" refers to an increased or upregulated or a decreased or downregulated expression as detected by absence, presence, or at least two-fold change in the amount of transcribed messenger RNA or translated protein in a sample.

"Disorder" refers to conditions or diseases of the colon, including colon cancer and precancerous conditions such as premalignant polyps.

"Fragment" refers to a chain of consecutive nucleotides from about 200 to about 700 base pairs in length. Fragments may be used in PCR or hybridization technologies to identify related nucleic acid molecules and in binding assays to screen for a ligand. Nucleic acids and their ligands identified in this manner are useful as therapeutics to regulate replication, transcription or translation.

A "hybridization complex" is formed between a cDNA and a nucleic acid of a sample when the purines of one molecule hydrogen bond with the pyrimidines of the complementary molecule, e.g., 5'-A-G-T-C-3' base pairs with 3'-T-C-A-G-5'. The degree of complementarity and the use of

nucleotide analogs affect the efficiency and stringency of hybridization reactions.

"Identity" as applied to sequences, refers to the quantification (usually percentage) of nucleotide or residue matches between at least two sequences aligned using a standardized algorithm such as Smith-Waterman alignment (Smith and Waterman (1981) J Mol Biol 147:195-197), CLUSTALW (Thompson et al. (1994) Nucleic Acids Res 22:4673-4680), or BLAST2 (Altschul et al. (1997) supra). BLAST2 may be used in a standardized and reproducible way to insert gaps in one of the sequences in order to optimize alignment and to achieve a more meaningful comparison between them. "Similarity" as applied to proteins uses the same algorithms but takes into account conservative substitutions of nucleotides or residues.

"Ligand" refers to any agent, molecule, or compound which will bind specifically to a complementary site on a cDNA molecule or polynucleotide, or to an epitope or a protein. Such ligands stabilize or modulate the activity of polynucleotides or proteins and may be composed of inorganic or organic substances including nucleic acids, proteins, carbohydrates, fats, and lipids.

"Oligonucleotide" refers a single stranded molecule from about 18 to about 60 nucleotides in length which may be used in hybridization or amplification technologies or in regulation of replication, transcription or translation. Substantially equivalent terms are amplimer, primer, and oligomer.

"Portion" refers to any part of a protein used for any purpose which retains at least one biological or antigenic characteristic of a native protein; but especially, to an epitope for the screening of ligands or for the production of antibodies.

"Post-translational modification" of a protein can involve lipidation, glycosylation, phosphorylation, acetylation, racemization, proteolytic cleavage, and the like. These processes may occur synthetically or biochemically. Biochemical modifications will vary by cellular location, cell type, pH, enzymatic milieu, and the like.

"Probe" refers to a cDNA that hybridizes to at least one nucleic acid molecule in a sample. Where targets are single stranded, probes are complementary single strands. Probes can be labeled with reporter molecules for use in hybridization reactions including Southern, northern, in situ, dot blot, array, and like technologies or in screening assays.

"Protein" refers to a polypeptide or any portion thereof. An "oligopeptide" is an amino acid sequence from about five residues to about 15 residues that is used as part of a fusion protein to produce an antibody.

"Purified" refers to any molecule or compound that is separated from its natural environment and is preferably 60% free, and more preferably 90% free from other components with which it is naturally associated.

"Sample" is used in its broadest sense as containing nucleic acids, proteins, antibodies, and the like. A sample may comprise a bodily fluid; the soluble fraction of a cell preparation, or an aliquot of

media in which cells were grown; a chromosome, an organelle, or membrane isolated or extracted from a cell; genomic DNA, RNA, or cDNA in solution or bound to a substrate; a cell; a tissue or tissue biopsy; a tissue print; a fingerprint, buccal cells, skin, or hair; and the like.

"Specific binding" refers to a special and precise interaction between two molecules which is dependent upon their structure, particularly their molecular side groups. For example, the intercalation of a regulatory protein into the major groove of a DNA molecule, the hydrogen bonding along the backbone between two single stranded nucleic acids, or the binding between an epitope of a protein and an agonist, antagonist, or antibody.

"Substrate" refers to any rigid or semi-rigid support to which cDNAs or proteins are bound and includes membranes, filters, chips, slides, wafers, fibers, magnetic or nonmagnetic beads, gels, capillaries or other tubing, plates, polymers, and microparticles with a variety of surface forms including wells, trenches, pins, channels and pores.

"Variant" refers to molecules that are recognized variations of a cDNA or a protein encoded by the cDNA. Splice variants may be determined by BLAST score, wherein the score is at least 100, and most preferably at least 400. Allelic variants have a high percent identity to the cDNAs and may differ by about three bases per hundred bases. "Single nucleotide polymorphism" (SNP) refers to a change in a single base as a result of a substitution, insertion or deletion. The change may be conservative (purine for purine) or non-conservative (purine to pyrimidine) and may or may not result in a change in an encoded amino acid.

The Invention

The present invention provides for a combination comprising a plurality of cDNAs or their complements, SEQ ID NOs:1-3, 5, 6, 8-10,12, 14, 15, 17, 18, 20, 22, 24, 26-29, 31, 33, 34, 36-39, 41-43, 45-47, 49, 51, 53. 55-58, 60, 62, 64, 66, 67, 69, 71, 72, 74-79, 81, 83-86, 88, 89, 91, 92, 94, 96, 97, 99, 100, 102-104, 106, 107, 109, 111, 112, 114, 116, 118, 119, 121, 123-126, 128, 130, 131-137, 139, 140, 142-151, 153-157, 159, 160, 162-165, 167-172, 174, 176, 177, 179-181, 183-187, 189-191, and 193 which may be used on a substrate to diagnose, to stage, to treat or to monitor the progression or treatment of colon cancer. These cDNAs represent known and novel genes differentially expressed in colon polyps and colon cancer. SEQ ID NOs:12, 41, 71, 74, 154,162, 167, 170, and 177 represent novel cDNAs associated with colon cancer. Since the novel cDNAs were identified solely by their differential expression, it is not essential to know a priori the name, structure, or function of the gene or it's encoded protein. The usefulness of the novel cDNAs exists in their immediate value as diagnostics for colon cancer.

Table 1 shows cDNA clones on an array having at least a 2-fold increase (upregulated) or decrease (downregulated, indicated by a minus sign) in at least 50% of the samples tested from patients with either colon polyps or colon cancer compared with normal colon. Column 1 shows the

Incyte Clone ID and columns 2-4 show differential expression values from patients with colon polyps, while columns 5-11 show values from patients with colon cancer. These genes are useful in diagnosing a precancerous condition in colon, or the presence of colon cancer.

Table 2 shows cDNA clones on an array that are differentially expressed in colon cancer relative to colon polyps. Column 1 shows the Incyte Clone ID and columns 2-4 show the differential expression values from patients with colon polyps, while columns 5-9 show the values from patients with colon cancer. Column 10 shows the P value for a t-test comparing colon polyps with colon tumor samples. Clones were selected on the basis of a P value in the t-test of less than or equal to 0.05, indicating a confidence level of at least 95% that the gene was differentially expressed in colon tumors to a greater or lesser extent than in colon polyps. These genes are useful in diagnosing colon cancer or monitoring the progression of a colon disorder from premalignant colon polyps to colon cancer.

Tables 3 and 4 further link the differentially expressed cDNA clones to full-length genes and to proteins in the Incyte database, and Table 5 provides the annotation of these sequences to known proteins in GenBank. Tables 6 and 7 provide further identification of encoded protein sequences by Pfam and the presence of signal peptide or transmembrane regions. Of particular note in Table 5 is SEQ ID NO:72, Incyte Template ID 1808144CB1 that is identified as a human mucosa associated mRNA (DRA; down-regulated in adenoma), g291963, a gene known to be down-regulated in colon adenomas and adenocarcinomas.

The cDNAs of the invention define a differential expression pattern for colon cancer or for a premalignant condition leading to colon cancer. Experimentally, differential expression of the cDNAs can be evaluated by methods including, but not limited to, differential display by spatial immobilization or by gel electrophoresis, genome mismatch scanning, representational discriminant analysis, clustering, transcript imaging and array technologies. These methods may be used alone or in combination.

The combination may be arranged on a substrate and hybridized with tissues from subjects with diagnosed colon disorders to identify those sequences which are differentially expressed in either colon cancer or premalignant colon polyps. This allows identification of those sequences of highest diagnostic and potential therapeutic value. In one embodiment, an additional set of cDNAs, such as cDNAs encoding signaling molecules, are arranged on the substrate with the combination. Such combinations may be useful in the elucidation of pathways which are affected in a particular disorder or to identify new, coexpressed, candidate, therapeutic molecules.

In another embodiment, the combination can be used for large scale genetic or gene expression analysis of a large number of novel, nucleic acid molecules. These samples are prepared by methods well known in the art and are from mammalian cells or tissues which are in a certain

stage of development; have been treated with a known molecule or compound, such as a cytokine, growth factor, a drug, and the like; or have been extracted or biopsied from a mammal with a known or unknown condition, disorder, or disease before or after treatment. The sample nucleic acid molecules are hybridized to the combination for the purpose of defining a novel gene profile associated with that developmental stage, treatment, or disorder.

cDNAs and Their Uses

cDNAs can be prepared by a variety of synthetic or enzymatic methods well known in the art. cDNAs can be synthesized, in whole or in part, using chemical methods well known in the art (Caruthers et al. (1980) Nucleic Acids Symp Ser (7)215-233). Alternatively, cDNAs can be produced enzymatically or recombinantly, by in vitro or in vivo transcription.

Nucleotide analogs can be incorporated into cDNAs by methods well known in the art. The only requirement is that the incorporated analog must base pair with native purines or pyrimidines. For example, 2, 6-diaminopurine can substitute for adenine and form stronger bonds with thymidine than those between adenine and thymidine. A weaker pair is formed when hypoxanthine is substituted for guanine and base pairs with cytosine. Additionally, cDNAs can include nucleotides that have been derivatized chemically or enzymatically.

cDNAs can be synthesized on a substrate. Synthesis on the surface of a substrate may be accomplished using a chemical coupling procedure and a piezoelectric printing apparatus as described by Baldeschweiler et al. (PCT publication WO95/251116). Alternatively, the cDNAs can be synthesized on a substrate surface using a self-addressable electronic device that controls when reagents are added as described by Heller et al. (USPN 5,605,662). cDNAs can be synthesized directly on a substrate by sequentially dispensing reagents for their synthesis on the substrate surface or by dispensing preformed DNA fragments to the substrate surface. Typical dispensers include a micropipette delivering solution to the substrate with a robotic system to control the position of the micropipette with respect to the substrate. There can be a multiplicity of dispensers so that reagents can be delivered to the reaction regions efficiently.

cDNAs can be immobilized on a substrate by covalent means such as by chemical bonding procedures or UV irradiation. In one method, a cDNA is bound to a glass surface which has been modified to contain epoxide or aldehyde groups. In another method, a cDNA is placed on a polylysine coated surface and UV cross-linked to it as described by Shalon et al. (WO95/35505). In yet another method, a cDNA is actively transported from a solution to a given position on a substrate by electrical means (Heller, supra). cDNAs do not have to be directly bound to the substrate, but rather can be bound to the substrate through a linker group. The linker groups are typically about 6 to 50 atoms long to provide exposure of the attached cDNA. Preferred linker groups include ethylene glycol oligomers, diamines, diacids and the like. Reactive groups on the substrate surface react with a terminal group

of the linker to bind the linker to the substrate. The other terminus of the linker is then bound to the cDNA. Alternatively, polynucleotides, plasmids or cells can be arranged on a filter. In the latter case, cells are lysed, proteins and cellular components degraded, and the DNA is coupled to the filter by UV cross-linking.

The cDNAs may be used for a variety of purposes. For example, the combination of the invention may be used on an array. The array, in turn, can be used in high-throughput methods for detecting a related polynucleotide in a sample, screening a plurality of molecules or compounds to identify a ligand, diagnosing a colon cancer, or inhibiting or inactivating a therapeutically relevant gene related to the cDNA.

When the cDNAs of the invention are employed on a microarray, the cDNAs are arranged in an ordered fashion so that each cDNA is present at a specified location. Because the cDNAs are at specified locations on the substrate, the hybridization patterns and intensities, which together create a unique expression profile, can be interpreted in terms of expression levels of particular genes and can be correlated with a particular metabolic process, condition, disorder, disease, stage of disease, or treatment.

Hybridization

The cDNAs or fragments or complements thereof may be used in various hybridization technologies. The cDNAs may be labeled using a variety of reporter molecules by either PCR recombinant, or enzymatic techniques. For example, a commercially available vector containing the cDNA is transcribed in the presence of an appropriate polymerase, such as T7 or SP6 polymerase, and at least one labeled nucleotide. Commercial kits are available for labeling and cleanup of such cDNAs. Radioactive (Amersham Pharmacia Biotech (APB), Piscataway NJ), fluorescent (Operon Technologies, Alameda CA), and chemiluminescent labeling (Promega, Madison WI) are well known in the art.

A cDNA may represent the complete coding region of an mRNA or be designed or derived from unique regions of the mRNA or genomic molecule, an intron, a 3' untranslated region, or from a conserved motif. The cDNA is at least 18 contiguous nucleotides in length and is usually single stranded. Such a cDNA may be used under hybridization conditions that allow binding only to an identical sequence, a naturally occurring molecule encoding the same protein, or an allelic variant. Discovery of related human and mammalian sequences may also be accomplished using a pool of degenerate cDNAs and appropriate hybridization conditions. Generally, a cDNA for use in Southern or northern hybridizations may be from about 400 to about 6000 nucleotides long. Such cDNAs have high binding specificity in solution-based or substrate-based hybridizations. An oligonucleotide, a fragment of the cDNA, may be used to detect a polynucleotide in a sample using PCR.

The stringency of hybridization is determined by G+C content of the cDNA, salt

concentration, and temperature. In particular, stringency is increased by reducing the concentration of salt or raising the hybridization temperature. In solutions used for some membrane based hybridizations, addition of an organic solvent such as formamide allows the reaction to occur at a lower temperature. Hybridization may be performed with buffers, such as 5x saline sodium citrate (SSC) with 1% sodium dodecyl sulfate (SDS) at 60°C, that permit the formation of a hybridization complex between nucleic acid sequences that contain some mismatches. Subsequent washes are performed with buffers such as 0.2xSSC with 0.1% SDS at either 45°C (medium stringency) or 65°-68°C (high stringency). At high stringency, hybridization complexes will remain stable only where the nucleic acid molecules are completely complementary. In some membrane-based hybridizations, preferably 35% or most preferably 50%, formamide may be added to the hybridization solution to reduce the temperature at which hybridization is performed. Background signals may be reduced by the use of detergents such as Sarkosyl or Triton X-100 (Sigma Aldrich, St. Louis MO) and a blocking agent such as denatured salmon sperm DNA. Selection of components and conditions for hybridization are well known to those skilled in the art and are reviewed in Ausubel et al. (1997, Short Protocols in Molecular Biology, John Wiley & Sons, New York NY, Units 2.8-2.11, 3.18-3.19 and 4.6-4.9).

Dot-blot, slot-blot, low density and high density arrays are prepared and analyzed using methods known in the art. cDNAs from about 18 consecutive nucleotides to about 5000 consecutive nucleotides in length are contemplated by the invention and used in array technologies. The preferred number of cDNAs on an array is at least about 100,000, a more preferred number is at least about 40,000, an even more preferred number is at least about 10,000, and a most preferred number is at least about 600 to about 800. The array may be used to monitor the expression level of large numbers of genes simultaneously and to identify genetic variants, mutations, and SNPs. Such information may be used to determine gene function; to understand the genetic basis of a disorder; to diagnose a disorder; and to develop and monitor the activities of therapeutic agents being used to control or cure a disorder. (See, e.g., USPN 5,474,796; WO95/11995; WO95/35505; USPN 5,605,662; and USPN 5,958,342.)

Screening and Purification Assays

A cDNA may be used to screen a library or a plurality of molecules or compounds for a ligand which specifically binds the cDNA. Ligands may be DNA molecules, RNA molecules, peptide nucleic acid molecules, peptides, proteins such as transcription factors, promoters, enhancers, repressors, and other proteins that regulate replication, transcription, or translation of the polynucleotide in the biological system. The assay involves combining the cDNA or a fragment thereof with the molecules or compounds under conditions that allow specific binding and detecting the bound cDNA to identify at least one ligand that specifically binds the cDNA.

In one embodiment, the cDNA may be incubated with a library of isolated and purified molecules or compounds and binding activity determined by methods such as a gel-retardation assay (USPN 6,010,849) or a reticulocyte lysate transcriptional assay. In another embodiment, the cDNA may be incubated with nuclear extracts from biopsied and/or cultured cells and tissues. Specific binding between the cDNA and a molecule or compound in the nuclear extract is initially determined by gel shift assay. Protein binding may be confirmed by raising antibodies against the protein and adding the antibodies to the gel-retardation assay where specific binding will cause a supershift in the assay.

In another embodiment, the cDNA may be used to purify a molecule or compound using affinity chromatography methods well known in the art. In one embodiment, the cDNA is chemically reacted with cyanogen bromide groups on a polymeric resin or gel. Then a sample is passed over and reacts with or binds to the cDNA. The molecule or compound which is bound to the cDNA may be released from the cDNA by increasing the salt concentration of the flow-through medium and then collected.

The cDNA may be used to purify a ligand from a sample. A method for using a cDNA to purify a ligand would involve combining the cDNA or a fragment thereof with a sample under conditions to allow specific binding, recovering the bound cDNA, and using an appropriate agent to separate the cDNA from the purified ligand.

Protein Production and Uses

The full length cDNAs or fragment thereof may be used to produce purified proteins using recombinant DNA technologies described herein and taught in Ausubel et al. (supra; Units 16.1-16.62). One of the advantages of producing proteins by these procedures is the ability to obtain highly-enriched sources of the proteins thereby simplifying purification procedures.

The proteins may contain amino acid substitutions, deletions or insertions made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues involved. Such substitutions may be conservative in nature when the substituted residue has structural or chemical properties similar to the original residue (e.g., replacement of leucine with isoleucine or valine) or they may be nonconservative when the replacement residue is radically different (e.g., a glycine replaced by a tryptophan). Computer programs included in LASERGENE software (DNASTAR, Madison WI), MACVECTOR software (Genetics Computer Group, Madison WI) and RasMol software (Roger Sayle, University of Massachusetts, Amherst MA) may be used to help determine which and how many amino acid residues in a particular portion of the protein may be substituted, inserted, or deleted without abolishing biological or immunological activity.

Expression of Encoded Proteins

Expression of a particular cDNA may be accomplished by cloning the cDNA into a vector

and transforming this vector into a host cell. The cloning vector used for the construction of cDNA libraries in the LIFESEQ databases may also be used for expression. Such vectors usually contain a promoter and a polylinker useful for cloning, priming, and transcription. An exemplary vector may also contain the promoter for ß-galactosidase, an amino-terminal methionine and the subsequent seven amino acid residues of ß-galactosidase. The vector may be transformed into competent E. coli cells. Induction of the isolated bacterial strain with isopropylthiogalactoside (IPTG) using standard methods will produce a fusion protein that contains an N terminal methionine, the first seven residues of ß-galactosidase, about 15 residues of linker, and the protein encoded by the cDNA.

The cDNA may be shuttled into other vectors known to be useful for expression of protein in specific hosts. Oligonucleotides containing cloning sites and fragments of DNA sufficient to hybridize to stretches at both ends of the cDNA may be chemically synthesized by standard methods. These primers may then be used to amplify the desired fragments by PCR. The fragments may be digested with appropriate restriction enzymes under standard conditions and isolated using gel electrophoresis. Alternatively, similar fragments are produced by digestion of the cDNA with appropriate restriction enzymes and filled in with chemically synthesized oligonucleotides. Fragments of the coding sequence from more than one gene may be ligated together and expressed.

Signal sequences that dictate secretion of soluble proteins are particularly desirable as component parts of a recombinant sequence. For example, a chimeric protein may be expressed that includes one or more additional purification-facilitating domains. Such domains include, but are not limited to, metal-chelating domains that allow purification on immobilized metals, protein A domains that allow purification on immobilized immunoglobulin, and the domain utilized in the FLAGS extension/affinity purification system (Immunex, Seattle WA). The inclusion of a cleavable-linker sequence such as ENTEROKINASEMAX (Invitrogen, San Diego CA) between the protein and the purification domain may also be used to recover the protein.

Suitable host cells may include, but are not limited to, mammalian cells such as Chinese Hamster Ovary (CHO) and human 293 cells, insect cells such as Sf9 cells, plant cells such as Nicotiana tabacum, yeast cells such as Saccharomyces cerevisiae, and bacteria such as E. coli. For each of these cell systems, a useful vector may also include an origin of replication and one or two selectable markers to allow selection in bacteria as well as in a transformed eukaryotic host. Vectors for use in eukaryotic host cells may require the addition of 3' poly(A) tail if the cDNA lacks poly(A).

Additionally, the vector may contain promoters or enhancers that increase gene expression. Many promoters are known and used in the art. Most promoters are host specific and exemplary promoters include SV40 promoters for CHO cells; T7 promoters for bacterial hosts; viral promoters and enhancers for plant cells; and PGH promoters for yeast. Adenoviral vectors with the rous sarcoma virus enhancer or retroviral vectors with long terminal repeat promoters may be used to drive

protein expression in mammalian cell lines. Once homogeneous cultures of recombinant cells are obtained, large quantities of secreted soluble protein may be recovered from the conditioned medium and analyzed using chromatographic methods well known in the art. An alternative method for the production of large amounts of secreted protein involves the transformation of mammalian embryos and the recovery of the recombinant protein from milk produced by transgenic cows, goats, sheep, and the like.

In addition to recombinant production, proteins or portions thereof may be produced manually, using solid-phase techniques (Stewart et al. (1969) Solid-Phase Peptide Synthesis, WH Freeman, San Francisco CA; Merrifield (1963) J Am Chem Soc 5:2149-2154), or using machines such as the ABI 431A peptide synthesizer (Applied Biosystems, Foster City CA). Proteins produced by any of the above methods may be used as pharmaceutical compositions to treat disorders associated with null or inadequate expression of the genomic sequence.

Screening and Purification Assays

A protein or a portion thereof encoded by the cDNA may be used to screen a library or a plurality of molecules or compounds for a ligand with specific binding affinity or to purify a molecule or compound from a sample. The protein or portion thereof employed in such screening may be free in solution, affixed to an abiotic or biotic substrate, or located intracellularly. For example, viable or fixed prokaryotic host cells that are stably transformed with recombinant nucleic acids that have expressed and positioned a protein on their cell surface can be used in screening assays. The cells are screened against a library or a plurality of ligands and the specificity of binding or formation of complexes between the expressed protein and the ligand may be measured. The ligands may be DNA, RNA, or PNA molecules, agonists, antagonists, antibodies, immunoglobulins, inhibitors, peptides, pharmaceutical agents, proteins, drugs, or any other test molecule or compound that specifically binds the protein. An exemplary assay involves combining the mammalian protein or a portion thereof with the molecules or compounds under conditions that allow specific binding and detecting the bound protein to identify at least one ligand that specifically binds the protein.

This invention also contemplates the use of competitive drug screening assays in which neutralizing antibodies capable of binding the protein specifically compete with a test compound capable of binding to the protein or oligopeptide or fragment thereof. One method for high throughput screening using very small assay volumes and very small amounts of test compound is described in USPN 5,876,946. Molecules or compounds identified by screening may be used in a model system to evaluate their toxicity, diagnostic, or therapeutic potential.

The protein may be used to purify a ligand from a sample. A method for using a protein to purify a ligand would involve combining the protein or a portion thereof with a sample under conditions to allow specific binding, recovering the bound protein, and using an appropriate chaotropic agent to

separate the protein from the purified ligand.

Production of Antibodies

A protein encoded by a cDNA of the invention may be used to produce specific antibodies. Antibodies may be produced using an oligopeptide or a portion of the protein with inherent immunological activity. Methods for producing antibodies include: 1) injecting an animal, usually goats, rabbits, or mice, with the protein, or an antigenically-effective portion or an oligopeptide thereof, to induce an immune response; 2) engineering hybridomas to produce monoclonal antibodies; 3) inducing in vivo production in the lymphocyte population; or 4) screening libraries of recombinant immunoglobulins. Recombinant immunoglobulins may be produced as taught in USPN 4,816,567.

Antibodies produced using the proteins of the invention are useful for the diagnosis of prepathologic disorders as well as the diagnosis of chronic or acute diseases characterized by abnormalities in the expression, amount, or distribution of the protein. A variety of protocols for competitive binding or immunoradiometric assays using either polyclonal or monoclonal antibodies specific for proteins are well known in the art. Immunoassays typically involve the formation of complexes between a protein and its specific binding molecule or compound and the measurement of complex formation. Immunoassays may employ a two-site, monoclonal-based assay that utilizes monoclonal antibodies reactive to two noninterfering epitopes on a specific protein or a competitive binding assay (Pound (1998) Immunochemical Protocols, Humana Press, Totowa NJ).

Immunoassay procedures may be used to quantify expression of the protein in cell cultures, in subjects with a particular disorder or in model animal systems under various conditions. Increased or decreased production of proteins as monitored by immunoassay may contribute to knowledge of the cellular activities associated with developmental pathways, engineered conditions or diseases, or treatment efficacy. The quantity of a given protein in a given tissue may be determined by performing immunoassays on freeze-thawed detergent extracts of biological samples and comparing the slope of the binding curves to binding curves generated by purified protein.

Labeling of Molecules for Assay

A wide variety of reporter molecules and conjugation techniques are known by those skilled in the art and may be used in various cDNA, polynucleotide, protein, peptide or antibody assays. Synthesis of labeled molecules may be achieved using commercial kits for incorporation of a labeled nucleotide such as ^{32}P-dCTP, Cy3-dCTP or Cy5-dCTP or amino acid such as ^{35}S-methionine. Polynucleotides, cDNAs, proteins, or antibodies may be directly labeled with a reporter molecule by chemical conjugation to amines, thiols and other groups present in the molecules using reagents such as BIODIPY or FITC (Molecular Probes, Eugene OR).

The proteins and antibodies may be labeled for purposes of assay by joining them, either covalently or noncovalently, with a reporter molecule that provides for a detectable signal. A wide

variety of labels and conjugation techniques are known and have been reported in the scientific and patent literature including, but not limited to USPN 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149; and 4,366,241.

DIAGNOSTICS

The cDNAs, or fragments thereof, may be used to detect and quantify differential gene expression; absence, presence, or excess expression of mRNAs; or to monitor mRNA levels during therapeutic intervention. Disorders associated with altered or differential expression include colon cancer and premalignant colon polyps. These cDNAs can also be utilized as markers of treatment efficacy against the disorders noted above and other disorders, conditions, and diseases over a period ranging from several days to months. The diagnostic assay may use hybridization or amplification technology to compare gene expression in a biological sample from a patient to standard samples in order to detect altered or differential gene expression. Qualitative or quantitative methods for this comparison are well known in the art.

For example, the cDNA may be labeled by standard methods and added to a biological sample from a patient under conditions for hybridization complex formation. After an incubation period, the sample is washed and the amount of label (or signal) associated with hybridization complexes is quantified and compared with a standard value. If the amount of label in the patient sample is significantly altered in comparison to the standard value, then the presence of the associated condition, disease or disorder is indicated.

In order to provide a basis for the diagnosis of a condition, disease or disorder associated with gene expression, a normal or standard expression profile is established. This may be accomplished by combining a biological sample taken from normal subjects, either animal or human, with a probe under conditions for hybridization or amplification. Standard hybridization may be quantified by comparing the values obtained using normal subjects with values from an experiment in which a known amount of a substantially purified target sequence is used. Standard values obtained in this manner may be compared with values obtained from samples from patients who are symptomatic for a particular condition, disease, or disorder. Deviation from standard values toward those associated with a particular condition is used to diagnose that condition.

Such assays may also be used to evaluate the efficacy of a particular therapeutic treatment regimen in animal studies and in clinical trial or to monitor the treatment of an individual patient. Once the presence of a condition is established and a treatment protocol is initiated, diagnostic assays may be repeated on a regular basis to determine if the level of expression in the patient begins to approximate that which is observed in a normal subject. The results obtained from successive assays may be used to show the efficacy of treatment over a period ranging from several days to months.

Gene Expression Profiles

A gene expression profile comprises a plurality of cDNAs and a plurality of detectable hybridization complexes, wherein each complex is formed by hybridization of one or more probes to one or more complementary sequences in a sample. The cDNAs of the invention are used as elements on a microarray to analyze gene expression profiles. In one embodiment, the microarray is used to monitor the progression of disease. Researchers can assess and catalog the differences in gene expression between healthy and diseased tissues or cells. By analyzing changes in patterns of gene expression, disease can be diagnosed at earlier stages before the patient is symptomatic. The invention can be used to formulate a prognosis and to design a treatment regimen. The invention can also be used to monitor the efficacy of treatment. For treatments with known side effects, the microarray is employed to improve the treatment regimen. A dosage is established that causes a change in genetic expression patterns indicative of successful treatment. Expression patterns associated with the onset of undesirable side effects are avoided. This approach may be more sensitive and rapid than waiting for the patient to show inadequate improvement, or to manifest side effects, before altering the course of treatment.

In another embodiment, animal models which mimic a human disease can be used to characterize expression profiles associated with a particular condition, disorder or disease; or treatment of the condition, disorder or disease. Novel treatment regimens may be tested in these animal models using microarrays to establish and then follow expression profiles over time. In addition, microarrays may be used with cell cultures or tissues removed from animal models to rapidly screen large numbers of candidate drug molecules, looking for ones that produce an expression profile similar to those of known therapeutic drugs, with the expectation that molecules with the same expression profile will likely have similar therapeutic effects. Thus, the invention provides the means to rapidly determine the molecular mode of action of a drug.

Assays Using Antibodies

Antibodies directed against epitopes on a protein encoded by a cDNA of the invention may be used in assays to quantify the amount of protein found in a particular human cell. Such assays include methods utilizing the antibody and a label to detect expression level under normal or disease conditions. The antibodies may be used with or without modification, and labeled by joining them, either covalently or noncovalently, with a labeling moiety.

Protocols for detecting and measuring protein expression using either polyclonal or monoclonal antibodies are well known in the art. Examples include ELISA, RIA, and fluorescent activated cell sorting (FACS). Such immunoassays typically involve the formation of complexes between the protein and its specific antibody and the measurement of such complexes. These and other assays are described in Pound (supra). The method may employ a two-site, monoclonal-based immunoassay utilizing monoclonal antibodies reactive to two non-interfering epitopes, or a competitive

binding assay. (See, e.g., Coligan et al. (1997) Current Protocols in Immunology, Wiley-Interscience, New York NY; Pound, supra)

THERAPEUTICS

The cDNAs and fragments thereof can be used in gene therapy. cDNAs can be delivered ex vivo to target cells, such as cells of bone marrow. Once stable integration and transcription and or translation are confirmed, the bone marrow may be reintroduced into the subject. Expression of the protein encoded by the cDNA may correct a colon cancer or premalignant colon polyps associated with mutation of a normal sequence, reduction or loss of an endogenous target protein, or overepression of an endogenous or mutant protein. Alternatively, cDNAs may be delivered in vivo using vectors such as retrovirus, adenovirus, adeno-associated virus, herpes simplex virus, and bacterial plasmids. Non-viral methods of gene delivery include cationic liposomes, polylysine conjugates, artificial viral envelopes, and direct injection of DNA (Anderson (1998) Nature 392:25-30; Dachs et al. (1997) Oncol Res 9:313-325; Chu et al. (1998) J Mol Med 76(3-4):184-192; Weiss et al. (1999) Cell Mol Life Sci 55(3):334-358; Agrawal (1996) Antisense Therapeutics, Humana Press, Totowa NJ; and August et al. (1997) Gene Therapy (Advances in Pharmacology, Vol. 40), Academic Press, San Diego CA).

In addition, expression of a particular protein can be regulated through the specific binding of a fragment of a cDNA to a genomic sequence or an mRNA which encodes the protein or directs its transcription or translation. The cDNA can be modified or derivatized to any RNA-like or DNA-like material including peptide nucleic acids, branched nucleic acids, and the like. These sequences can be produced biologically by transforming an appropriate host cell with a vector containing the sequence of interest.

Molecules which regulate the activity of the cDNA or encoded protein are useful as therapeutics for colon cancer and premalignant colon polyps. Such molecules include agonists which increase the expression or activity of the polynucleotide or encoded protein, respectively; or antagonists which decrease expression or activity of the polynucleotide or encoded protein, respectively. In one aspect, an antibody which specifically binds the protein may be used directly as an antagonist or indirectly as a delivery mechanism for bringing a pharmaceutical agent to cells or tissues which express the protein.

Additionally, any of the proteins, or their ligands, or complementary nucleic acid sequences may be administered as pharmaceutical compositions or in combination with other appropriate therapeutic agents. Selection of the appropriate agents for use in combination therapy may be made by one of ordinary skill in the art, according to conventional pharmaceutical principles. The combination of therapeutic agents may act synergistically to affect the treatment or prevention of the conditions and disorders associated with an immune response. Using this approach, one may be able

to achieve therapeutic efficacy with lower dosages of each agent, thus reducing the potential for adverse side effects. Further, the therapeutic agents may be combined with pharmaceutically-acceptable carriers including excipients and auxiliaries which facilitate processing of the active compounds into preparations which can be used pharmaceutically. Further details on techniques for formulation and administration used by doctors and pharmacists may be found in the latest edition of Remington's Pharmaceutical Sciences (Mack Publishing, Easton PA).

Model Systems

Animal models may be used as bioassays where they exhibit a phenotypic response similar to that of humans and where exposure conditions are relevant to human exposures. Mammals are the most common models, and most infectious agent, cancer, drug, and toxicity studies are performed on rodents such as rats or mice because of low cost, availability, lifespan, reproductive potential, and abundant reference literature. Inbred and outbred rodent strains provide a convenient model for investigation of the physiological consequences of underexpression or overexpression of genes of interest and for the development of methods for diagnosis and treatment of diseases. A mammal inbred to overexpress a particular gene (for example, secreted in milk) may also serve as a convenient source of the protein expressed by that gene.

Transgenic Animal Models

Transgenic rodents that overexpress or underexpress a gene of interest may be inbred and used to model human diseases or to test therapeutic or toxic agents. (See, e.g., USPN 5,175,383 and USPN 5,767,337.) In some cases, the introduced gene may be activated at a specific time in a specific tissue type during fetal or postnatal development. Expression of the transgene is monitored by analysis of phenotype, of tissue-specific mRNA expression, or of serum and tissue protein levels in transgenic animals before, during, and after challenge with experimental drug therapies.

Embryonic Stem Cells

Embryonic (ES) stem cells isolated from rodent embryos retain the potential to form embryonic tissues. When ES cells such as the mouse 129/SvJ cell line are placed in a blastocyst from the C57BL/6 mouse strain, they resume normal development and contribute to tissues of the live-born animal. ES cells are preferred for use in the creation of experimental knockout and knockin animals. The method for this process is well known in the art and the steps are: the cDNA is introduced into a vector, the vector is transformed into ES cells, transformed cells are identified and microinjected into mouse cell blastocysts, blastocysts are surgically transferred to pseudopregnant dams. The resulting chimeric progeny are genotyped and bred to produce heterozygous or homozygous strains.

Knockout Analysis

In gene knockout analysis, a region of a gene is enzymatically modified to include a non-natural intervening sequence such as the neomycin phosphotransferase gene (neo; Capecchi (1989)

Science 244:1288-1292). The modified gene is transformed into cultured ES cells and integrates into the endogenous genome by homologous recombination. The inserted sequence disrupts transcription and translation of the endogenous gene.

Knockin Analysis

ES cells can be used to create knockin humanized animals or transgenic animal models of human diseases. With knockin technology, a region of a human gene is injected into animal ES cells, and the human sequence integrates into the animal cell genome. Transgenic progeny or inbred lines are studied and treated with potential pharmaceutical agents to obtain information on the progression and treatment of the analogous human condition.

As described herein, the uses of the cDNAs, provided in the Sequence Listing of this application, and their encoded proteins are exemplary of known techniques and are not intended to reflect any limitation on their use in any technique that would be known to the person of average skill in the art. Furthermore, the cDNAs provided in this application may be used in molecular biology techniques that have not yet been developed, provided the new techniques rely on properties of nucleotide sequences that are currently known to the person of ordinary skill in the art, e.g., the triplet genetic code, specific base pair interactions, and the like. Likewise, reference to a method may include combining more than one method for obtaining or assembling full length cDNA sequences that will be known to those skilled in the art. It is also to be understood that this invention is not limited to the particular methodology, protocols, and reagents described, as these may vary. It is also understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the present invention which will be limited only by the appended claims. The examples below are provided to illustrate the subject invention and are not included for the purpose of limiting the invention.

EXAMPLES

I Construction of cDNA Libraries

RNA was purchased from Clontech Laboratories (Palo Alto CA) or isolated from various tissues. Some tissues were homogenized and lysed in guanidinium isothiocyanate, while others were homogenized and lysed in phenol or in a suitable mixture of denaturants, such as TRIZOL reagent (Life Technologies, Rockville MD). The resulting lysates were centrifuged over CsCl cushions or extracted with chloroform. RNA was precipitated with either isopropanol or ethanol and sodium acetate, or by other routine methods.

Phenol extraction and precipitation of RNA were repeated as necessary to increase RNA purity. In most cases, RNA was treated with DNAse. For most libraries, poly(A) RNA was isolated using oligo d(T)-coupled paramagnetic particles (Promega), OLIGOTEX latex particles (Qiagen, Valencia CA), or an OLIGOTEX mRNA purification kit (Qiagen). Alternatively, poly(A) RNA was

isolated directly from tissue lysates using other kits, including the POLY(A)PURE mRNA purification kit (Ambion, Austin TX).

In some cases, Stratagene (La Jolla CA) was provided with RNA and constructed the corresponding cDNA libraries. Otherwise, cDNA was synthesized and cDNA libraries were constructed with the UNIZAP vector system (Stratagene) or SUPERSCRIPT plasmid system (Life Technologies) using the recommended procedures or similar methods known in the art. (See Ausubel, supra, Units 5.1 through 6.6.) Reverse transcription was initiated using oligo d(T) or random primers. Synthetic oligonucleotide adapters were ligated to double stranded cDNA, and the cDNA was digested with the appropriate restriction enzyme or enzymes. For most libraries, the cDNA was size-selected (300-1000 bp) using SEPHACRYL S1000, SEPHAROSE CL2B, or SEPHAROSE CL4B column chromatography (APB) or preparative agarose gel electrophoresis. cDNAs were ligated into compatible restriction enzyme sites of the polylinker of the pBLUESCRIPT phagemid (Stratagene), pSPORT1 plasmid (Life Technologies), or pINCY plasmid (Incyte Genomics, Inc., Palo Alto CA). Recombinant plasmids were transformed into XL1-BLUE, XL1-BLUEMRF, or SOLR competent E. coli cells (Stratagene) or DH5α, DH10B, or ELECTROMAX DH10B competent E. coli cells (Life Technologies).

In some cases, libraries were superinfected with a 5x excess of the helper phage, M13K07, according to the method of Vieira et al. (1987, Methods Enzymol 153:3-11) and normalized or subtracted using a methodology adapted from Soares (1994, Proc Natl Acad Sci 91:9228-9232), Swaroop et al. (1991, Nucleic Acids Res 19:1954), and Bonaldo et al. (1996, Genome Research 6:791-806). The modified Soares normalization procedure was utilized to reduce the repetitive cloning of highly expressed high abundance cDNAs while maintaining the overall sequence complexity of the library. Modification included significantly longer hybridization times which allowed for increased gene discovery rates by biasing the normalized libraries toward those infrequently expressed low-abundance cDNAs which are poorly represented in a standard transcript image (Soares et al., supra).

II Isolation and Sequencing of cDNA Clones

Plasmids were recovered from host cells by in vivo excision using the UNIZAP vector system (Stratagene) or by cell lysis. Plasmids were purified using one of the following: the Magic or WIZARD MINIPREPS DNA purification system (Promega); the AGTC MINIPREP purification kit (Edge BioSystems, Gaithersburg MD); the QIAWELL 8, QIAWELL 8 Plus, or QIAWELL 8 Ultra plasmid purification systems, or the REAL PREP 96 plasmid purification kit (QIAGEN). Following precipitation, plasmids were resuspended in 0.1 ml of distilled water and stored, with or without lyophilization, at 4°C.

Alternatively, plasmid DNA was amplified from host cell lysates using direct link PCR in a high-throughput format (Rao (1994) Anal Biochem 216:1-14). Host cell lysis and thermal cycling

steps were carried out in a single reaction mixture. Samples were processed and stored in 384-well plates, and the concentration of amplified plasmid DNA was quantified fluorometrically using PICOGREEN dye (Molecular Probes) and a FLUOROSKAN II fluorescence scanner (Labsystems Oy, Helsinki Finland).

cDNA sequencing reactions were processed using standard methods or high-throughput instrumentation such as the ABI CATALYST 800 thermal cycler (Applied Biosystems) or the DNA ENGINE thermal cycler (MJ Research, Watertown MA) in conjunction with the HYDRA microdispenser (Robbins Scientific, Sunnyvale CA) or the MICROLAB 2200 system (Hamilton, Reno NV). cDNA sequencing reactions were prepared using reagents provided by APB or supplied in ABI sequencing kits such as the ABI PRISM BIGDYE cycle sequencing kit (Applied Biosystems). Electrophoretic separation of cDNA sequencing reactions and detection of labeled cDNAs were carried out using the MEGABACE 1000 DNA sequencing system (APB); the ABI PRISM 373 or 377 sequencing systems (Applied Biosystems) in conjunction with standard ABI protocols and base calling software; or other sequence analysis systems known in the art. Reading frames within the cDNA sequences were identified using standard methods (reviewed in Ausubel, supra, Unit 7.7).

III Extension of cDNA Sequences

Nucleic acid sequences were extended using the cDNA clones and oligonucleotide primers. One primer was synthesized to initiate 5' extension of the known fragment, and the other, to initiate 3' extension of the known fragment. The initial primers were designed using OLIGO 4.06 software (National Biosciences, Plymouth MN), or another appropriate program, to be about 22 to 30 nucleotides in length, to have a GC content of about 50% or more, and to anneal to the target sequence at temperatures of about 68°C to about 72°C. Any stretch of nucleotides which would result in hairpin structures and primer-primer dimerizations was avoided.

Selected human cDNA libraries were used to extend the sequence. If more than one extension was necessary or desired, additional or nested sets of primers were designed. Preferred libraries are ones that have been size-selected to include larger cDNAs. Also, random primed libraries are preferred because they will contain more sequences with the 5' and upstream regions of genes. A randomly primed library is particularly useful if an oligo d(T) library does not yield a full-length cDNA.

High fidelity amplification was obtained by PCR using methods well known in the art. PCR was performed in 96-well plates using the DNA ENGINE thermal cycler (MJ Research). The reaction mix contained DNA template, 200 nmol of each primer, reaction buffer containing Mg^{2+}, $(NH_4)_2SO_4$, and β-mercaptoethanol, Taq DNA polymerase (APB), ELONGASE enzyme (Life Technologies), and Pfu DNA polymerase (Stratagene), with the following parameters for primer pair

PCI A and PCI B (Incyte Genomics): Step 1: 94°C, 3 min; Step 2: 94°C, 15 sec; Step 3: 60°C, 1 min; Step 4: 68°C, 2 min; Step 5: Steps 2, 3, and 4 repeated 20 times; Step 6: 68°C, 5 min; Step 7: storage at 4°C. In the alternative, the parameters for primer pair T7 and SK+ (Stratagene) were as follows: Step 1: 94°C, 3 min; Step 2: 94°C, 15 sec; Step 3: 57°C, 1 min; Step 4: 68°C, 2 min; Step 5: Steps 2, 3, and 4 repeated 20 times; Step 6: 68°C, 5 min; Step 7: storage at 4°C.

The concentration of DNA in each well was determined by dispensing 100 µl PICOGREEN reagent (0.25% reagent in 1x TE, v/v; Molecular Probes) and 0.5 µl of undiluted PCR product into each well of an opaque fluorimeter plate (Corning Costar, Acton MA) and allowing the DNA to bind to the reagent. The plate was scanned in a FLUOROSKAN II (Labsystems Oy) to measure the fluorescence of the sample and to quantify the concentration of DNA. A 5 µl to 10 µl aliquot of the reaction mixture was analyzed by electrophoresis on a 1% agarose mini-gel to determine which reactions were successful in extending the sequence.

The extended nucleic acids were desalted and concentrated, transferred to 384-well plates, digested with CviJI cholera virus endonuclease (Molecular Biology Research, Madison WI), and sonicated or sheared prior to religation into pUC18 vector (APB). For shotgun sequencing, the digested nucleic acids were separated on low concentration (0.6 to 0.8%) agarose gels, fragments were excised, and agar digested with AGARACE enzyme (Promega). Extended clones were religated using T4 DNA ligase (New England Biolabs, Beverly MA) into pUC18 vector (APB), treated with Pfu DNA polymerase (Stratagene) to fill-in restriction site overhangs, and transformed into competent E. coli cells. Transformed cells were selected on antibiotic-containing media, and individual colonies were picked and cultured overnight at 37°C in 384-well plates in LB/2x carbenicillin liquid media.

The cells were lysed, and DNA was amplified by PCR using Taq DNA polymerase (APB) and Pfu DNA polymerase (Stratagene) with the following parameters: Step 1: 94°C, 3 min; Step 2: 94°C, 15 sec; Step 3: 60°C, 1 min; Step 4: 72°C, 2 min; Step 5: steps 2, 3, and 4 repeated 29 times; Step 6: 72°C, 5 min; Step 7: storage at 4°C. DNA was quantified using PICOGREEN reagent (Molecular Probes) as described above. Samples with low DNA recoveries were reamplified using the same conditions described above. Samples were diluted with 20% dimethylsulfoxide (DMSO; 1:2, v/v), and sequenced using DYENAMIC energy transfer sequencing primers and the DYENAMIC DIRECT cycle sequencing kit (APB) or the ABI PRISM BIGDYE terminator cycle sequencing kit (Applied Biosystems).

IV Assembly and Analysis of Sequences

Component nucleotide sequences from chromatograms were subjected to PHRED analysis (Phil Green, University of Washington, Seattle WA) and assigned a quality score. The sequences having at least a required quality score were subject to various pre-processing algorithms to eliminate

low quality 3' ends, vector and linker sequences, polyA tails, Alu repeats, mitochondrial and ribosomal sequences, bacterial contamination sequences, and sequences smaller than 50 base pairs. Sequences were screened using the BLOCK 2 program (Incyte Genomics), a motif analysis program based on sequence information contained in the SWISS-PROT and PROSITE databases (Bairoch et al. (1997) Nucleic Acids Res 25:217-221; Attwood et al. (1997) J Chem Inf Comput Sci 37:417-424).

Processed sequences were subjected to assembly procedures in which the sequences were assigned to bins, one sequence per bin. Sequences in each bin were assembled to produce consensus sequences, templates. Subsequent new sequences were added to existing bins using BLAST (Altschul (supra); Altschul et al. (supra); Karlin et al. (1988) Proc Natl Acad Sci 85:841-845), BLASTn (vers.1.4, WashU), and CROSSMATCH software (Phil Green, supra). Candidate pairs were identified as all BLAST hits having a quality score greater than or equal to 150. Alignments of at least 82% local identity were accepted into the bin. The component sequences from each bin were assembled using PHRAP (Phil Green, supra). Bins with several overlapping component sequences were assembled using DEEP PHRAP (Phil Green, supra).

Bins were compared against each other, and those having local similarity of at least 82% were combined and reassembled. Reassembled bins having templates of insufficient overlap (less than 95% local identity) were re-split. Assembled templates were also subjected to analysis by STITCHER/EXON MAPPER algorithms which analyzed the probabilities of the presence of splice variants, alternatively spliced exons, splice junctions, differential expression of alternative spliced genes across tissue types, disease states, and the like. These resulting bins were subjected to several rounds of the above assembly procedures to generate the template sequences found in the LIFESEQ GOLD database (Incyte Genomics).

The assembled templates were annotated using the following procedure. Template sequences were analyzed using BLASTn (vers. 2.0, NCBI) versus GBpri (GenBank version 117). "Hits" were defined as an exact match having from 95% local identity over 200 base pairs through 100% local identity over 100 base pairs, or a homolog match having an E-value equal to or greater than $1x10^{-8}$. (The "E-value" quantifies the statistical probability that a match between two sequences occurred by chance). The hits were subjected to frameshift FASTx versus GENPEPT (GenBank version 117). In this analysis, a homolog match was defined as having an E-value of $1x10^{-8}$. The assembly method used above was described in USSN 09/276,534, filed March 25, 1999, and the LIFESEQ GOLD user manual (Incyte Genomics).

Following assembly, template sequences were subjected to motif, BLAST, Hidden Markov Model (HMM; Pearson and Lipman (1988) Proc Natl Acad Sci 85:2444-2448; Smith and Waterman (1981) J Mol Biol 147:195-197), and functional analyses, and categorized in protein hierarchies using methods described in USSN 08/812,290, filed March 6, 1997; USSN 08/947,845, filed October 9,

1997; USPN 5,953,727; and USSN 09/034,807, filed March 4, 1998. Template sequences may be further queried against public databases such as the GenBank rodent, mammalian, vertebrate, eukaryote, prokaryote, and human EST databases.

V Selection of Sequences, Microarray Preparation and Use

Incyte clones represent template sequences derived from the LIFESEQ GOLD assembled human sequence database (Incyte Genomics). In cases where more than one clone was available for a particular template, the 5'-most clone in the template was used on the microarray. The HUMAN GENOME GEM series 1-3 microarrays (Incyte Genomics) contain 28,626 array elements which represent 10,068 annotated clusters and 18,558 unannotated clusters. For the UNIGEM series microarrays (Incyte Genomics), Incyte clones were mapped to non-redundant Unigene clusters (Unigene database (build 46), NCBI; Shuler (1997) J Mol Med 75:694-698), and the 5' clone with the strongest BLAST alignment (at least 90% identity and 100 bp overlap) was chosen, verified, and used in the construction of the microarray. The UNIGEM V microarray (Incyte Genomics) contains 7075 array elements which represent 4610 annotated genes and 2,184 unannotated clusters. Tables 1 and 2 show the GenBank annotations for SEQ ID NOs:1-138 of this invention as produced by BLAST analysis.

To construct microarrays, cDNAs were amplified from bacterial cells using primers complementary to vector sequences flanking the cDNA insert. Thirty cycles of PCR increased the initial quantity of cDNAs from 1-2 ng to a final quantity of greater than 5 μg. Amplified cDNAs were then purified using SEPHACRYL-400 columns (APB). Purified cDNAs were immobilized on polymer-coated glass slides. Glass microscope slides (Corning, Corning NY) were cleaned by ultrasound in 0.1% SDS and acetone, with extensive distilled water washes between and after treatments. Glass slides were etched in 4% hydrofluoric acid (VWR Scientific Products, West Chester PA), washed thoroughly in distilled water, and coated with 0.05% aminopropyl silane (Sigma Aldrich) in 95% ethanol. Coated slides were cured in a 110°C oven. cDNAs were applied to the coated glass substrate using a procedure described in USPN 5,807,522. One microliter of the cDNA at an average concentration of 100 ng/μl was loaded into the open capillary printing element by a high-speed robotic apparatus which then deposited about 5 nl of cDNA per slide.

Microarrays were UV-crosslinked using a STRATALINKER UV-crosslinker (Stratagene), and then washed at room temperature once in 0.2% SDS and three times in distilled water. Non-specific binding sites were blocked by incubation of microarrays in 0.2% casein in phosphate buffered saline (Tropix, Bedford MA) for 30 minutes at 60°C followed by washes in 0.2% SDS and distilled water as before.

VI Preparation of Samples

Tissue Samples

Matched normal colon and cancerous colon or colon polyp tissue samples were provided by the Huntsman Cancer Institute, (Salt Lake City, UT). Donor 3754 is an individual diagnosed with a pendunculated colon polyp; age and sex of the donor is unknown. Donor 3755 is an individual diagnosed with colon polyps and having a family history of colon cancer; age and sex of the donor is unknown. Donor 3583 is a 58 year-old male diagnosed with a tubulovillous adenoma hyperplastic polyp. Donor 3311 is an 85 year-old male diagnosed with an invasive, poorly differentiated adenocarcinoma with metastases to the lymph nodes. Donor 3756 is a 78 year-old female diagnosed with an invasive, moderately differentiated adenocarcinoma. Donor 3757 is a 75 year-old female diagnosed with an invasive, moderate to poorly differentiated adenocarcinoma with metastases to the lymph nodes. Donor 3649 is an 86 year-old individual, sex unknown, diagnosed with an invasive, well-differentiated adenocarcinoma. Donor 3647 is an 83 year-old individual, sex unknown, diagnosed with an invasive, moderately well-differentiated adenocarcinoma with metastases to the lymph nodes. Donor 3839 is a 60 year-old individual, sex unknown, diagnosed with colon cancer. Donor 3581 is a male of unknown age diagnosed with a colorectal tumor. Donors 3754, 3755, 3311, 3756, and 3757 were matched against a common control sample comprising a pool of normal colon tissue from three additional donors. All other comparisons were done with matched normal and tumor or polyp tissue from the same donor.

Isolation and Labeling of Sample cDNAs

Tissues were homogenized and lysed in 1 ml of TRIZOL reagent (5 x 10^6 cells/ml; Life Technologies). The lysates were vortexed thoroughly and incubated at room temperature for 2-3 minutes and extracted with 0.5 ml chloroform. The extract was mixed, incubated at room temperature for 5 minutes, and centrifuged at 15,000 rpm for 15 minutes at 4°C. The aqueous layer was collected and an equal volume of isopropanol was added. Samples were mixed, incubated at room temperature for 10 minutes, and centrifuged at 15,000 rpm for 20 minutes at 4°C. The supernatant was removed and the RNA pellet was washed with 1 ml of 70% ethanol, centrifuged at 15,000 rpm at 4°C, and resuspended in RNAse-free water. The concentration of the RNA was determined by measuring the optical density at 260 nm.

Poly(A) RNA was prepared using an OLIGOTEX mRNA kit (QIAGEN) with the following modifications: OLIGOTEX beads were washed in tubes instead of on spin columns, resuspended in elution buffer, and then loaded onto spin columns to recover mRNA. To obtain maximum yield, the mRNA was eluted twice.

Each poly(A) RNA sample was reverse transcribed using MMLV reverse-transcriptase, 0.05 pg/µl oligo-d(T) primer (21mer), 1x first strand buffer, 0.03 units/ul RNase inhibitor, 500 uM dATP, 500 uM dGTP, 500 uM dTTP, 40 uM dCTP, and 40 uM either dCTP-Cy3 or dCTP-Cy5 (APB). The

reverse transcription reaction was performed in a 25 ml volume containing 200 ng poly(A) RNA using the GEMBRIGHT kit (Incyte Genomics). Specific control poly(A) RNAs (YCFR06, YCFR45, YCFR67, YCFR85, YCFR43, YCFR22, YCFR23, YCFR25, YCFR44, YCFR26) were synthesized by in vitro transcription from non-coding yeast genomic DNA (W. Lei, unpublished). As quantitative controls, control mRNAs (YCFR06, YCFR45, YCFR67, and YCFR85) at 0.002ng, 0.02ng, 0.2 ng, and 2ng were diluted into reverse transcription reaction at ratios of 1:100,000, 1:10,000, 1:1000, 1:100 (w/w) to sample mRNA, respectively. To sample differential expression patterns, control mRNAs (YCFR43, YCFR22, YCFR23, YCFR25, YCFR44, YCFR26) were diluted into reverse transcription reaction at ratios of 1:3, 3:1, 1:10, 10:1, 1:25, 25:1 (w/w) to sample mRNA. Reactions were incubated at 37°C for 2 hr, treated with 2.5 ml of 0.5M sodium hydroxide, and incubated for 20 minutes at 85°C to the stop the reaction and degrade the RNA.

cDNAs were purified using two successive CHROMA SPIN 30 gel filtration spin columns (Clontech). Cy3- and Cy5-labeled reaction samples were combined as described below and ethanol precipitated using 1 ml of glycogen (1 mg/ml), 60 ml sodium acetate, and 300 ml of 100% ethanol. The cDNAs were then dried to completion using a SpeedVAC system (Savant Instruments, Holbrook NY) and resuspended in 14 µl 5x SSC, 0.2% SDS.

VII Hybridization and Detection

Hybridization reactions contained 9 µl of sample mixture containing 0.2 µg each of Cy3 and Cy5 labeled cDNA synthesis products in 5x SSC, 0.2% SDS hybridization buffer. The mixture was heated to 65°C for 5 minutes and was aliquoted onto the microarray surface and covered with an 1.8 cm^2 coverslip. The microarrays were transferred to a waterproof chamber having a cavity just slightly larger than a microscope slide. The chamber was kept at 100% humidity internally by the addition of 140 µl of 5x SSC in a corner of the chamber. The chamber containing the microarrays was incubated for about 6.5 hours at 60°C. The microarrays were washed for 10 min at 45°C in low stringency wash buffer (1x SSC, 0.1% SDS), three times for 10 minutes each at 45°C in high stringency wash buffer (0.1x SSC), and dried.

Reporter-labeled hybridization complexes were detected with a microscope equipped with an Innova 70 mixed gas 10 W laser (Coherent, Santa Clara CA) capable of generating spectral lines at 488 nm for excitation of Cy3 and at 632 nm for excitation of Cy5. The excitation laser light was focused on the microarray using a 20x microscope objective (Nikon, Melville NY). The slide containing the microarray was placed on a computer-controlled X-Y stage on the microscope and raster-scanned past the objective. The 1.8 cm x 1.8 cm microarray used in the present example was scanned with a resolution of 20 micrometers.

In two separate scans, the mixed gas multiline laser excited the two fluorophores sequentially. Emitted light was split, based on wavelength, into two photomultiplier tube detectors (PMT R1477;

Hamamatsu Photonics Systems, Bridgewater NJ) corresponding to the two fluorophores. Appropriate filters positioned between the microarray and the photomultiplier tubes were used to filter the signals. The emission maxima of the fluorophores used were 565 nm for Cy3 and 650 nm for Cy5. Each microarray was typically scanned twice, one scan per fluorophore using the appropriate filters at the laser source, although the apparatus was capable of recording the spectra from both fluorophores simultaneously.

The sensitivity of the scans was calibrated using the signal intensity generated by a cDNA control species. Samples of the calibrating cDNA were separately labeled with the two fluorophores and identical amounts of each were added to the hybridization mixture. A specific location on the microarray contained a complementary DNA sequence, allowing the intensity of the signal at that location to be correlated with a weight ratio of hybridizing species of 1:100,000.

The output of the photomultiplier tube was digitized using a 12-bit RTI-835H analog-to-digital (A/D) conversion board (Analog Devices, Norwood, MA) installed in an IBM-compatible PC computer. The digitized data were displayed as an image where the signal intensity was mapped using a linear 20-color transformation to a pseudocolor scale ranging from blue (low signal) to red (high signal). The data was also analyzed quantitatively. Where two different fluorophores were excited and measured simultaneously, the data were first corrected for optical crosstalk (due to overlapping emission spectra) between the fluorophores using each fluorophore's emission spectrum.

A grid was superimposed over the fluorescence signal image such that the signal from each spot was centered in each element of the grid. The fluorescence signal within each element was then integrated to obtain a numerical value corresponding to the average intensity of the signal. The software used for signal analysis was the GEMTOOLS gene expression analysis program (Incyte Genomics). Significance was defined as signal to background ratio exceeding 2x and area hybridization exceeding 40%.

VIII Data Analysis and Results

Array elements that exhibited at least 2-fold change in expression, a signal intensity over 250 units, a signal-to-background ratio of at least 2.5, and an element spot size of at least 40% were identified as differentially expressed using the GEMTOOLS program (Incyte Genomics). Differential expression values were converted to log base 2 scale. The cDNAs that are differentially expressed are shown in Tables 1 and 2. The cDNAs identified in Table 1 are differentially expressed at least 2-fold in at least 50% of patient samples tested. These genes are useful diagnostic markers or as potential therapeutic targets for premalignant colon polyps or colon cancer. The cDNAs identified in Table 2 showed a statistically greater differential expression pattern in colon cancer than colon polyps by t-test analysis. These genes are useful diagnostic markers for colon tumor progression from premalignant colon polyps to cancer or as potential therapeutic targets for colon cancer.

IX Other Hybridization Technologies and Analyses

Other hybridization technologies utilize a variety of substrates such as nylon membranes, capillary tubes, etc. Arranging cDNAs on polymer coated slides is described in Example V; sample cDNA preparation and hybridization and analysis using polymer coated slides is described in examples VI and VII, respectively.

The cDNAs are applied to a membrane substrate by one of the following methods. A mixture of cDNAs is fractionated by gel electrophoresis and transferred to a nylon membrane by capillary transfer. Alternatively, the cDNAs are individually ligated to a vector and inserted into bacterial host cells to form a library. The cDNAs are then arranged on a substrate by one of the following methods. In the first method, bacterial cells containing individual clones are robotically picked and arranged on a nylon membrane. The membrane is placed on LB agar containing selective agent (carbenicillin, kanamycin, ampicillin, or chloramphenicol depending on the vector used) and incubated at 37°C for 16 hr. The membrane is removed from the agar and consecutively placed colony side up in 10% SDS, denaturing solution (1.5 M NaCl, 0.5 M NaOH), neutralizing solution (1.5 M NaCl, 1 M Tris, pH 8.0), and twice in 2xSSC for 10 min each. The membrane is then UV irradiated in a STRATALINKER UV-crosslinker (Stratagene).

In the second method, cDNAs are amplified from bacterial vectors by thirty cycles of PCR using primers complementary to vector sequences flanking the insert. PCR amplification increases a starting concentration of 1-2 ng nucleic acid to a final quantity greater than 5 µg. Amplified nucleic acids from about 400 bp to about 5000 bp in length are purified using SEPHACRYL-400 beads (APB). Purified nucleic acids are arranged on a nylon membrane manually or using a dot/slot blotting manifold and suction device and are immobilized by denaturation, neutralization, and UV irradiation as described above.

Hybridization probes derived from cDNAs of the Sequence Listing are employed for screening cDNAs, mRNAs, or genomic DNA in membrane-based hybridizations. Probes are prepared by diluting the cDNAs to a concentration of 40-50 ng in 45 µl TE buffer, denaturing by heating to 100°C for five min and briefly centrifuging. The denatured cDNA is then added to a REDIPRIME tube (APB), gently mixed until blue color is evenly distributed, and briefly centrifuged. Five microliters of [^{32}P]dCTP is added to the tube, and the contents are incubated at 37°C for 10 min. The labeling reaction is stopped by adding 5 µl of 0.2M EDTA, and probe is purified from unincorporated nucleotides using a PROBEQUANT G-50 microcolumn (APB). The purified probe is heated to 100°C for five min and then snap cooled for two min on ice.

Membranes are pre-hybridized in hybridization solution containing 1% Sarkosyl and 1x high phosphate buffer (0.5 M NaCl, 0.1 M Na_2HPO_4, 5 mM EDTA, pH 7) at 55°C for two hr. The probe, diluted in 15 ml fresh hybridization solution, is then added to the membrane. The membrane is

hybridized with the probe at 55°C for 16 hr. Following hybridization, the membrane is washed for 15 min at 25°C in 1mM Tris (pH 8.0), 1% Sarkosyl, and four times for 15 min each at 25°C in 1mM Tris (pH 8.0). To detect hybridization complexes, XOMAT-AR film (Eastman Kodak, Rochester NY) is exposed to the membrane overnight at -70°C, developed, and examined.

X Further Characterization of Differentially Expressed cDNAs and Proteins

Clones were blasted against the LIFESEQ Gold 5.1 database (Incyte Genomics) and an Incyte template and its sequence variants were chosen for each clone. The template and variant sequences were blasted against GenBank database to acquire annotation. The nucleotide sequences were translated into amino acid sequence which was blasted against the GenPept and other protein databases to acquire annotation and characterization, i.e., structural motifs.

Percent sequence identity can be determined electronically for two or more amino acid or nucleic acid sequences using the MEGALIGN program (DNASTAR). The percent identity between two amino acid sequences is calculated by dividing the length of sequence A, minus the number of gap residues in sequence A, minus the number of gap residues in sequence B, into the sum of the residue matches between sequence A and sequence B, times one hundred. Gaps of low or of no homology between the two amino acid sequences are not included in determining percentage identity.

Sequences with conserved protein motifs may be searched using the BLOCKS search program. This program analyses sequence information contained in the Swiss-Prot and PROSITE databases and is useful for determining the classification of uncharacterized proteins translated from genomic or cDNA sequences (Bairoch et al.,supra; Attwood et al., supra). PROSITE database is a useful source for identifying functional or structural domains that are not detected using motifs due to extreme sequence divergence. Using weight matrices, these domains are calibrated against the SWISS-PROT database to obtain a measure of the chance distribution of the matches.

The PRINTS database can be searched using the BLIMPS search program to obtain protein family "fingerprints". The PRINTS database complements the PROSITE database by exploiting groups of conserved motifs within sequence alignments to build characteristic signatures of different protein families. For both BLOCKS and PRINTS analyses, the cutoff scores for local similarity were: >1300=strong, 1000-1300=suggestive; for global similarity were: p<exp-3; and for strength (degree of correlation) were: >1300=strong, 1000-1300=weak.

XI Expression of the Encoded Protein

Expression and purification of a protein encoded by a cDNA of the invention is achieved using bacterial or virus-based expression systems. For expression in bacteria, cDNA is subcloned into a vector containing an antibiotic resistance gene and an inducible promoter that directs high levels of cDNA transcription. Examples of such promoters include, but are not limited to, the *trp-lac* (*tac*)

hybrid promoter and the T5 or T7 bacteriophage promoter in conjunction with the *lac* operator regulatory element. Recombinant vectors are transformed into bacterial hosts, such as BL21(DE3). Antibiotic resistant bacteria express the protein upon induction with IPTG. Expression in eukaryotic cells is achieved by infecting Spodoptera frugiperda (Sf9) insect cells with recombinant baculovirus, Autographica californica nuclear polyhedrosis virus. The polyhedrin gene of baculovirus is replaced with the cDNA by either homologous recombination or bacterial-mediated transposition involving transfer plasmid intermediates. Viral infectivity is maintained and the strong polyhedrin promoter drives high levels of transcription.

For ease of purification, the protein is synthesized as a fusion protein with glutathione-S-transferase (GST; APB) or a similar alternative such as FLAG. The fusion protein is purified on immobilized glutathione under conditions that maintain protein activity and antigenicity. After purification, the GST moiety is proteolytically cleaved from the protein with thrombin. A fusion protein with FLAG, an 8-amino acid peptide, is purified using commercially available monoclonal and polyclonal anti-FLAG antibodies (Eastman Kodak, Rochester NY).

XII Production of Specific Antibodies

A denatured protein from a reverse phase HPLC separation is obtained in quantities up to 75 mg. This denatured protein is used to immunize mice or rabbits following standard protocols. About 100 μg is used to immunize a mouse, while up to 1 mg is used to immunize a rabbit. The denatured protein is radioiodinated and incubated with murine B-cell hybridomas to screen for monoclonal antibodies. About 20 mg of protein is sufficient for labeling and screening several thousand clones.

In another approach, the amino acid sequence translated from a cDNA of the invention is analyzed using PROTEAN software (DNASTAR) to determine regions of high antigenicity, essentially antigenically-effective epitopes of the protein. The optimal sequences for immunization are usually at the C-terminus, the N-terminus, and those intervening, hydrophilic regions of the protein that are likely to be exposed to the external environment when the protein is in its natural conformation. Typically, oligopeptides about 15 residues in length are synthesized using an ABI 431 peptide synthesizer (Applied Biosystems) using Fmoc-chemistry and then coupled to keyhole limpet hemocyanin (KLH; Sigma Aldrich) by reaction with M-maleimidobenzoyl-N-hydroxysuccinimide ester. If necessary, a cysteine may be introduced at the N-terminus of the peptide to permit coupling to KLH. Rabbits are immunized with the oligopeptide-KLH complex in complete Freund's adjuvant. The resulting antisera are tested for antipeptide activity by binding the peptide to plastic, blocking with 1% BSA, reacting with rabbit antisera, washing, and reacting with radioiodinated goat anti-rabbit IgG.

Hybridomas are prepared and screened using standard techniques. Hybridomas of interest are detected by screening with radioiodinated protein to identify those fusions producing a monoclonal antibody specific for the protein. In a typical protocol, wells of 96 well plates (FAST,

Becton-Dickinson, Palo Alto CA) are coated with affinity-purified, specific rabbit-anti-mouse (or suitable anti-species Ig) antibodies at 10 mg/ml. The coated wells are blocked with 1% BSA and washed and exposed to supernatants from hybridomas. After incubation, the wells are exposed to radiolabeled protein at 1 mg/ml. Clones producing antibodies bind a quantity of labeled protein that is detectable above background.

Such clones are expanded and subjected to 2 cycles of cloning at 1 cell/3 wells. Cloned hybridomas are injected into pristane-treated mice to produce ascites, and monoclonal antibody is purified from the ascitic fluid by affinity chromatography on protein A (APB). Monoclonal antibodies with affinities of at least 10^8 M^{-1}, preferably 10^9 to 10^{10} M^{-1} or stronger, are made by procedures well known in the art.

XIII Purification of Naturally Occurring Protein Using Specific Antibodies

Naturally occurring or recombinant protein is substantially purified by immunoaffinity chromatography using antibodies specific for the protein. An immunoaffinity column is constructed by covalently coupling the antibody to CNBr-activated SEPHAROSE resin (APB). Media containing the protein is passed over the immunoaffinity column, and the column is washed using high ionic strength buffers in the presence of detergent to allow preferential absorbance of the protein. After coupling, the protein is eluted from the column using a buffer of pH 2-3 or a high concentration of urea or thiocyanate ion to disrupt antibody/protein binding, and the protein is collected.

XIV Screening Molecules for Specific Binding with the cDNA or Protein

The cDNA or fragments thereof and the protein or portions thereof are labeled with ^{32}P-dCTP, Cy3-dCTP, Cy5-dCTP (APB), or BIODIPY or FITC (Molecular Probes), respectively. Candidate molecules or compounds previously arranged on a substrate are incubated in the presence of labeled nucleic or amino acid. After incubation under conditions for either a cDNA or a protein, the substrate is washed, and any position on the substrate retaining label, which indicates specific binding or complex formation, is assayed. The binding molecule is identified by its arrayed position on the substrate. Data obtained using different concentrations of the nucleic acid or protein are used to calculate affinity between the labeled nucleic acid or protein and the bound molecule. High throughput screening using very small assay volumes and very small amounts of test compound is fully described in Burbaum et al. USPN 5,876,946.

All patents and publications mentioned in the specification are incorporated herein by reference. Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it

should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the described modes for carrying out the invention that are obvious to those skilled in the field of molecular biology or related fields are intended to be within the scope of the following claims.

5

PCT/US2005/000000